

I see it in your eyes :
Learnings of the world’s shallowest CNN predicting
emotions from muted in-the-wild video-chats real-time

Vedhas Pandit¹, Maximilian Schmitt¹, Nicholas Cummins¹, Björn Schuller^{1,2}

Abstract

A robust value- and time-continuous emotion recognition is of huge relevance to healthcare. A real-time patient monitoring system understanding the patient’s physical and emotional state can help doctors make an appropriate diagnosis and treatment plan. In-person and online conversations with patients are beneficial likewise. Presence and absence of assistive technology, that is able to detect cues a human could miss, can be a difference between a life and a death.

For the first time ever, we present shallowest realisable CNN consisting of a single filter in its only intermediate 1D convolutional layer; learnings of which are proven to be fairly applicable to German, Hungarian and Chinese populations. We draw insights from visualisations of the trained filter-weights and the facial action unit (FAU) activations of the participants featured in SEWA’s in-the-wild, spontaneous video-chat sessions, i. e., the provided input data.

Indeed, the learnings of CNNs are consistent with the human perception of human emotion expression. A fairly decent crosscultural predictive ability of the trained CNNs is a testimony to universality of FAU features in understanding human emotional expressions. Because the FAU features can be extracted real-time using openFACE, and because the presented model is shallow, this study proposes a promising architecture for a robust, end-to-end, value- and time-continuous, pan-cultural real-time emotion recognition.

¹ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

²GLAM – Group on Language, Audio & Music, Imperial College London, UK

Keywords: Feature relevance, in the wild, real-time, emotion recognition, model explainability, layerwise relevance propagation, pattern attribution.

1. Introduction

Importance of social media environments as a new source of data for healthcare cannot be overstated today. The potential for the health care industry to leverage existing social media data for better quality, cost, and quality measurement is unparalleled. Likewise, deep learning advancements have made number-crunching and information extraction possible at a scale that is unprecedented. This simultaneous growth has opened doors to many next-generation applications that can utilise social media data and deep learning for a better healthcare.

1.1. Relevance of social media analytics to healthcare

Innovative approaches to building social data-driven health information help to create better healthcare decisions across the entire health care domain – for patients, providers, and health systems. Because the social media content gives us a sneak peek into the lives, emotions and opinions of the real people like never before, we can leverage this data to recognise patterns that relate to their health needs. We can gain insights into habits and preferences of the people, which treatments work the best for them and which ones don't – as a consequence of their habits and preferences, or otherwise (e.g., because of other factors such as ethnicity, age, and gender) []. Such information can help health professionals gain a much better understanding of the treatments and patient education that a population is most receptive to, and adopt accordingly []. Social media data are also valuable for business, who create the "marketing" and "social" elements of a social media strategy that are key to successful use of social media to improve engagement and engagement with customers and advertisers. Likewise, social data analytics is a powerful way to discover people that will likely lead a fitness regime, and to also recognise the precise stimulus to keep them motivated []. Social media enables patients to share their personal experience with a particular

product or a service (e.g., a hospital facility, doctors, treatment, medicines) to a much larger audience. This in turn can help healthcare professionals likewise to identify the gaps in the system, which in turn can, for example, help reduce the waiting times, improve the customer satisfaction, doctor-patient relationship, likely delivering also a better health outcome. The potential of the data to identify patient-specific treatment options is one area worth focusing on by health providers and patients alike.

1.2. Emotion-aware AI, and the unique challenges

Data mining, i.e., creating artificial intelligence based application has helped improve diagnosis and treatment of various diseases, e.g., diagnosis of cancer, diabetes, and cardiovascular disease and assessment of psychological health [1]. The technology, therefore, could be used to identify ‘red flag’ situations in real-time, which could then be avoided. For example, we could identify people prone to depression or suicide, or those who had negative and potentially life-threatening experiences in the past to help guide them through the challenges of living with a chronic illness - helping to make it an easier transition. However, such a system needs to meet very high standards of requirements. First, it needs to be robust enough to be used on non-laboratory, unconstrained, in-the-wild data featuring spontaneous behaviours. Secondly, it is desirable that it captures various nuances of emotions/affect, without resorting to a crude classification into three (positive, negative and neutral) or six basic emotion classes. Value-continuous multidimensional affect regression is preferable. It is desirable to be able to track one’s emotions continuously in time. The prediction model should ideally be also fast enough to be able to run in real-time. Fortunately, growth of machine intelligence or artificial intelligence (AI) have made all these goals within our reach.

The progress in the data mining technology, simultaneous to rise in social media content, has been stupendous. From convincingly mimicking the human speech right down to even individual voice [2], to generating high resolution ‘deep-fake’ videos [3], to penning down a highly plausible and coherent synthetic

text on any given topic [], the advancements in AI are staggering. Such is the growth that finesse with which AI performs these tasks remains hard to believe still today, while ironically, these capabilities and their usage across the board
60 is becoming a common knowledge day by day. For example, it might be hard to believe from the human-like coherency of the text that a few introductory paragraphs of this very manuscript have been partly generated using the very recently released GPT-2 language model []³, which learnt the English language on its own by simply browsing through the internet. While advancements of
65 such a scale are a boon to predictive power of this technology, these also present some unique challenges. It is reasonable to anticipate surge in synthesised social media content, consequently a corruption of social media content coming in from real humans. While stronger and smarter AI discriminators are being proposed, the generators too are getting more powerful day by day. The boundaries
70 between a synthesised and a real data are slowly beginning to disappear. Advancement in machine learning, particularly in the deep learning domain theory and frameworks, has been a primary driving factor to this progress.

It is because of the deep learning technology, that a real-time value- and time-continuous affect recognition is no longer a distant reality. The primary
75 and the earliest criticism of deep learning methodology (continuing up to even until very recent times) has long been its inexplicability. There are monumental developments in the ‘model explainability’ domain as well lately. Decision explainability is especially crucial for an application domain like healthcare, where people’s lives are at stake. In this paper, we unearth learnings of a shallow
80 CNN as it learns to predict human emotions from in-the-wild videos value- and time-continuously – without assistance of any audio or textual data – of people coming from very different cultures, contexts, gender and age-groups.

In Section 2, we discuss research so far that is directly relevant to our study. In the sections next, we discuss in detail our dataset (Section 3), experiments
85 (Section 4) and insights we gained from these experiments (Section 5). In Sec-

³<https://talktotransformer.com/>

tion 6, we summarise our findings with limitations of this study, and our planned future work in conclusion.

2. Related Research

Table 1: Participant count and duration for the data splits of the SEWA dataset

Culture	Partition	#Subjects	Duration
German	Training	34	93:12
German	Development	14	37:46
German	Test	16	46:38
Hungarian	Test	66	133:12
Chinese	Test	70	

3. Dataset

As discussed in the previous section, for the lack of as cross-cultural, and as in-the-wild dyadic conversational data, we use the SEWA dataset for this study. SEWA happens to be also a very popular dataset, available in the public domain for free, on which many benchmarking studies have been performed so far by various research groups from across the globe. This dataset has previously been used in various audiovisual emotion recognition challenges (e.g., AVEC 2017, AVEC 2018, AVEC 2019) [1]. The dataset features spontaneous dyadic conversations over the internet (i.e., video chat sessions) between participants, where they discuss an advertisement shown to them right before their discussions begin. The participants came from different age-groups, with various degrees of mutual acquaintance, different cultural backgrounds. They conversed in one of these six languages; Chinese, English, German, Greek, Hungarian, and Serbian. With 201 male and 197 female participants (male / female gender ratio = 1.020) and with about 60 pairs from each of the 6 linguistic populations, the the dataset features behavioural data from diversified and balanced demographics. So far, the audiovisual recordings of only Chinese, German and Hungarian participants have been fully annotated and are available for further research (cf. Section 3).

3.1. Labels and features in the dataset

SEWA dataset features value- and time-continuous annotations for arousal, valence and liking (how much participants liked the advertisement), annotated

110 by five to six annotators coming from the same linguistic background as of the participants of the audiovisuals they annotate. While annotating, therefore, the annotators can make use of every data modality available to them, namely the linguistic content (i.e., what is being spoken), the audio content (i.e., how it is being spoken, e.g., prosody, pitch, tone of the voice), and the visual content
 115 (i.e., non-verbal cues, facial expressions).

Correspondingly, to train a machine learning algorithm, in addition to audiovisual recordings and transcriptions of the conversations, the dataset contains various other linguistic, audio and video features, e.g., Mel Frequency Cepstral Coefficients (MFCCs), their 1st and 2nd derivatives, EGEMAPS audio feature set [] extracted using OPENSIMILE [], raw locations of the facial
 120 landmark points, facial action unit (FAU) features extracted using OPENFACE [].

In this study, we restrict our attention to employing FAU features, in conjunction with speaker-turn activation feature – which is indicative of at least
 125 one of the two participants speaking. We try to investigate how the neural networks utilise FAUs so efficiently, what makes their learnings generalise across different cultures, and if their learnings are consistent with our perception of human emotions.

















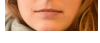

















3.2. Facial Action Units

130 An FAU is defined by Facial Action Coding System (FACS) – a predefined set of different simultaneous facial muscle movements. In the experiments next, we use 17 of these, along with the associated confidence scores that were extracted simultaneously using the OPENFACE toolkit. These are listed in Table 2.

3.3. Feature and Label Statistics

135 To explore the utility and challenges presented by the extracted FAU feature set, we first prepared statistical summary of the activation values of the individual features we would later train our models with. We train our models with visual features identical to those provided to the participants of AVEC 2018

Table 2: Illustrations of what the FAU activations look like [].⁴

Action Unit (AU)	Description	Example Movement	
		From	To
1	Inner Brow Raiser		
2	Outer Brow Raiser		
4	Brow Lowerer		
5	Upper Lid Raiser		
6	Cheek Raiser		
7	Lid Tightener		
9	Nose Wrinkler		
10	Upper Lip Raiser		
12	Lip Corner Puller		
14	Dimpler		
15	Lip Corner Depressor		
17	Chin Raiser		
20	Lip stretcher		
23	Lip Tightener		
25	Lips part		
26	Jaw Drop		
45	Blink		

and AVEC 2019 challenge; i.e., the moving average and the moving standard
140 deviations of FAU activations computed for consecutive 50 frames (1 second)
with a hop of 5 frames (100 ms). Section 3.3 presents grouped box-plots of the
corresponding moving mean and moving standard deviation values, for better
comparison of these features across cultures. As can be seen from the plot, the
ranges, quartiles, outlier distributions are similar across different cultures. Con-

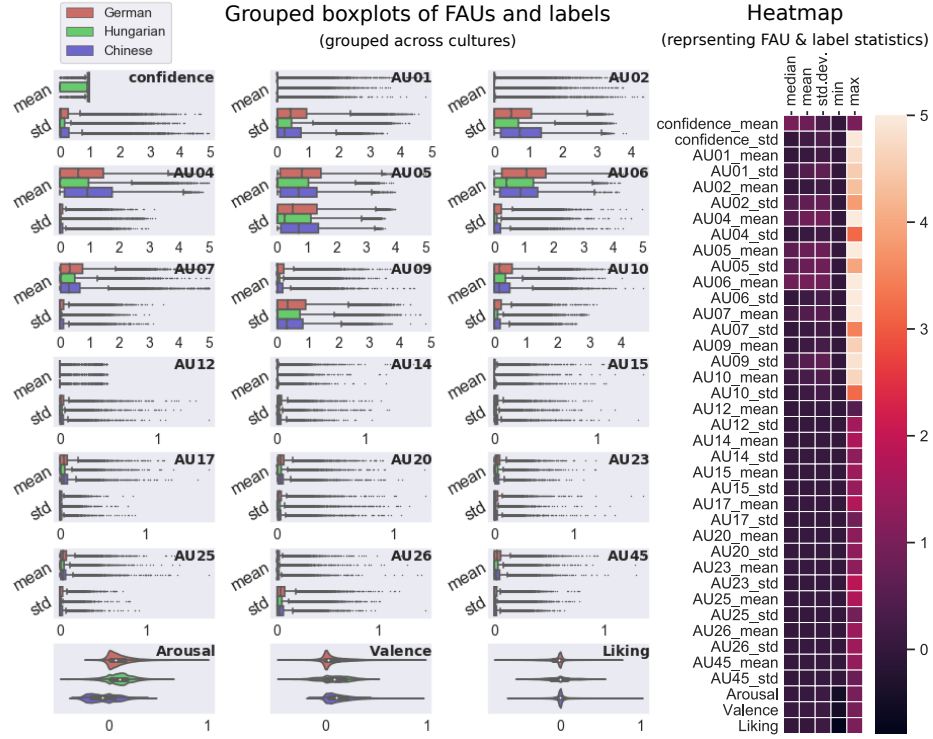


Figure 1: Boxplot of FAU feature activation levels and the arousal, valence, liking labels. As can be seen from the figure, the range (cf. x-axes) for different FAU features are drastically different. Consequently, a few FAU activations can dominate the rest of the features (cf. max values). On the other hand, a few other FAU features (e. g., AU12-mean) rarely ever manifest, even their maximum values are close to minimum of almost all other FAU feature activations.

145 sidering how different the three cultures are from one another, this consistency of the feature values across different culture – in some sense – is a testimony to universality of the FAU features. We can expect superior pan-cultural performance of the models trained using FAU features

Likely however, not all FAU features are equal. We notice that the range of values that the moving mean and standard deviations can take vary drastically across different FAUs. For example, mean AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10 computed from 1second of a video can be close to 5 maximally, while the rest of the FAU mean values hardly ever go higher than 1.5. The dissimilarity is so stark that even maximas of the certain feature values are

150

155 less than medians of most other features (cf. heatmap in Section 3.3). Such
extremely high or low values, in the absence of a batch-normalisation layer,
can drive prediction of a neural network astray. Interestingly enough, in spite of
the featured high maxims, medians of the moving means and moving standard
deviations for AU1 to AU10 features are always less than 0.5, in fact are mostly
160 close to 0.0.

4. Experiments

Taking inspiration from our previous work where we established that

5. Results

In this section, we present graphically some of the filters that the trained
165 CNNs used to predict nuanced arousal and valence values time-continuously.

6. Conclusion

References